AD0825512

AFFDL-TR-66-202 [redacted]

DESIGN OF SAMPLING PROCEDURES FOR FLIGHT LOADS STUDIES

S. C. Choi

Measurement Analysis Corporation

May 1967

This document is subject to special export controls and each transmittal to foreign governments or foreign nationals may be made only with prior approval of the Air Force Flight Dynamics Laboratory (FDTR), Wright-Patterson Air Force Base, Ohio 45433.

Air Force Flight Dynamics Laboratory
Research and Technology Division
Air Force System Command
Wright-Patterson Air Force Base, Ohio
45433

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Copies of this report should not be returned to the Research and Technology Division unless return is required by security considerations, contractual obligations, or notice on a specific document.

200 - June 1967 - CO192 - 36-883

AFFDL-TR-66-202

AD825512

DESIGN OF SAMPLING PROCEDURES FOR FLIGHT LOADS STUDIES

S. C. Choi

Measurement Analysis Corporation

FOREWORD

This report was prepared by the Measurement Analysis Corporation, Los Angeles, California, for the Theoretical Mechanics Branch, Structures Division, Air Force Flight Dynamics Laboratory, Wright-Patterson Air Force Base, Ohio, under Contract AF 33(615)-3123, Project No. BPSN 5(61136716-62405334). The research performed is part of a continuing effort to improve the evaluation and utilization of loads applied to the vehicle design and fatigue problem. Mr. E. Titus of the Theoretical Mechanics Branch, FDTR, was the Air Force project engineer.

This report covers work conducted from July 1965 through June 1966. Contractor's Report No. is MAC 510-03; the manuscript was released by the author June 1966 for publication as an AFFDL Technical Report.

The author gratefully expresses his appreciation for the assistance given by C. L. Gray of the Measurement Analysis Corporation and the cooperation of C. J. Schmid and E. Titus of the Theoretical Mechanics Branch.

This technical report has been reviewed and is approved.

Francis J. Janik, Jr.
Chief, Theoretical Mechanics Branch
Structures Division
Air Force Flight Dynamics Laboratory

ABSTRACT

The problems of estimating the total number of measurement points and the optimum spatial distribution of locations on a structure are approached theoretically in this report. The significant factors to be considered are statistical reliability and economy. Therefore, the relationships are developed with the emphasis on measurement efficiency. Random, systematic, and stratified sampling methods are compared for efficiency in estimating mean values. Then the optimum allocation of a fixed number of measurement points in stratified sampling is developed, and illustrative examples are given. Finally, relationships are presented which will allow the total sample size to be estimated under the assumptions of normal and log-normal sampling distributions as well as by a nonparametric approach. These formulas are deemed to be quite useful for experiment planning purposes.

This abstract is subject to special export controls and each transmittal to foreign governments or foreign nationals may be made only with prior approval of the Air Force Flight Dynamics Laboratory (FDTR), Wright-Patterson Air Force Base, Ohio 45433.

CONTENTS

GLOSSARY OF SYMBOLS

A	area of the structure
A_h	area of the structure in the hth zone
d	allowable measurement error
d'	allowable error in the log regime
L	number of nonoverlapping zones (disjoint strata)
n	total number of data points in a sample
n_h	number of data points in the hth zone
N	maximum sample size possible
P	probability that a sample value will exceed a critical value
P_h	probability that a sample value within the hth zone will exceed a critical value
s^2	sample variance
w	estimator for $\log \mu_X = \overline{\log X} + \frac{\sigma^2}{2}$
W_h	weighting function associated with the hth zone
X	random variable used here as mean square stress
X_i	ith sample value of X
X_c	established critical value for X
$\overline{X}$	sample mean of X
Y	derived random variable
Y_i	value of Y derived from the ith sample value of X
Z	random variable having a normal distribution
$Z_{\alpha/2}$	the $100\alpha/2$ percentage point of a standardized normal distribution

α	statistical level of significance
σ	true standard deviation
σ^2	true variance
μ	true mean
E[ ]	expected value of
Var[ ]	variance of
Prob[ ]	probability that
SE[ ]	standard error in
$(\hat{\ })$	estimated value of
$(\bar{\ })$	space averaged value of

1. INTRODUCTION

There are basically two statistical problems associated with sampling the response of aircraft structure to flight loads. Since these loads are typically representative of a time varying random process, the first item of concern is the time averaged statistical properties at a single point on the structure. Given sufficient sample record length, these properties can be estimated with reasonable accuracy, and for stationary random loading conditions, necessary record lengths are not difficult to obtain.

The second basic problem concerns point-to-point variation on a structure. Having evaluated conditions at a single point as a function of time, it is of interest to know how the remainder of the structure is behaving. It is this facet of sampling which is of concern in this report. The relationships between the number of data points in a structural sample and the accuracy in estimating two statistical properties of the data are discussed thoroughly. The two properties investigated are the mean value and the probability that a point selected at random on the structure will exhibit a sample value which exceeds some specified level. The developments are aimed at providing tools for the planning of statistical loads experiments. In Section 2, three different methods of sampling are described along with their relative efficiencies. Formulas are given in Section 3 for the optimum allocation of points in a sample when the structure has been partitioned into zones for study. Section 4 gives methods for determining the total number of points in a sample which will correspond to a required degree of accuracy in the estimates. This is done for the normal distribution, the log normal distribution, and a nonparametric approach which considers the distribution to be unknown.

2. SAMPLING METHODS

In this section, three methods of spatial sampling are discussed in terms of the parameters needed to assess their relative efficiencies. The example random variable being sampled is mean square stress, although any time averaged measure of load response could be substituted. It is assumed in all cases that the statistical uncertainty associated with finite sample record length is negligible compared with spatial variation. In the following, n denotes the number of data points in a sample and X_i the ith independent mean square stress measurement.

2.1 RANDOM SAMPLING

If mean square stress is measured at n points on a structure selected at random, one has a random sample of data. It is assumed that every point on the structure has been given independent and equal probability of being a data point. The mean value of the data in the sample thus gathered is given by

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{1}$$

The variance of the sample mean is defined as

$$\mathrm{Var}\,(\bar{X}) = \left(\frac{N-1}{N}\right) \frac{\sigma^2}{n} \tag{2}$$

where σ^2 is the variance based on the maximum possible sample size, N. Since N is usually considered infinite for structures, $(N - 1/N)$ is equal to one. For a sample size, n, the variance of the sample mean can be

estimated by

$$\widehat{\mathrm{Var}}\,(\bar{X}) = \frac{s_X^2}{n} \tag{3}$$

with $(\wedge)$ denoting estimated value. The term s_X^2, which estimates σ^2, is the unbiased sample variance and is given by

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{4}$$

To estimate the probability, P $(0 \leq P \leq 1)$, that the mean square stress at any point on the structure selected at random will exceed some critical value X_c, proceed as follows. Define a new random variable Y which has these properties. For any measurement in the sample, Y_i equals one if $X_i \geq X_c$ and is zero if $X_i < X_c$. Since the probability that $X_i \geq X_c$ is P, it follows that the expected value of each Y_i is

$$E[Y_i] = 1 \cdot P + 0(1 - P) = P \tag{5}$$

Then, if the random sample consists of n measurements, the estimate of P is given by

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{6}$$

For a variable such as Y which can assume only the values one and zero, the sample variance can be expressed as

$$s_Y^2 = \frac{n}{n-1}\,\hat{P}(1 - \hat{P}) \tag{7}$$

The variance of $\hat{P}$, using the relationship employed in Eq. (3) is then estimated by

$$\widehat{\mathrm{Var}}\,(\hat{P}) = \frac{1}{n-1}\,\hat{P}(1-\hat{P}) \tag{8}$$

2.2 SYSTEMATIC SAMPLING

The second method of sampling requires the sample points to be laid out on a structure in a systematic fashion. For example, in a one-dimensional structure, this may amount to taking measurements at uniform intervals along the single dimension. Systematic sampling, thus, places a restriction on sample point location, which was not the case for random sampling. One unfavorable aspect of this technique is the inability to correctly evaluate data when certain periodic trends exist. This problem is discussed in Reference 1 along with examples of special cases. In general, however, this sampling method should be avoided when periodic trends exist in the data. On the other hand, when a linear trend in the data exists, systematic sampling becomes much more efficient than random sampling. Periodic trends are more likely to exist in aircraft structural stress than are linear trends because of normal mode response.

The mean value of a systematic sample of mean square stress in a structure is given by Eq. (1), and the variance of the sample mean is closely approximated by

$$\mathrm{Var}\,(\bar{X}) = \frac{N-1}{N}\,\sigma^2 \tag{9}$$

when the variance due to finite time sampling is very small. Since the maximum possible sample size, N, can usually be considered infinite for structural sampling, (N - 1/N) is equal to one, and the variance of the

sample mean is simply σ^2. That is, it is independent of the sample size n. By comparing Eqs. (2) and (9), it is apparent that systematic sampling is, in general, less efficient than random sampling. In view of the above shortcomings of the method, systematic sampling will not be considered further in the discussion.

2.3 STRATIFIED SAMPLING

Using this method, a structure is partitioned into L non-overlapping zones (disjoint strata), and n measurement points are allocated over the structure in such a fashion that each zone contains at least two randomly located points. The sample mean of the measurements is computed from

$$\bar{X} = \sum_{h=1}^{L} W_h \bar{X}_h \tag{10}$$

where W_h is a weighting function associated with the hth zone, and $\bar{X}_h$ is the mean value for that zone. Here it can be seen that the interpretation of the weighting function, W_h, can be extremely important. Although many engineering factors may be involved in this interpretation, simple area relationships will be used in this development. Then, W_h will be defined as

$$W_h = \frac{A_h}{A} \tag{11}$$

where A_h is the area of the hth zone, and A is the total area of the structure under study. The variance of the sample mean is given by

$$\mathrm{Var}\,(\overline{X}) = \sum_{h=1}^{L} \frac{W_h^2\,\sigma_h^2}{n_h} \tag{12}$$

where n_h is the number of points in the hth zone, and σ_h^2 is the true zone variance. Variance of the sample mean is estimated from

$$\widehat{\mathrm{Var}}\,(\overline{X}) = \sum_{h=1}^{L} \frac{W_h^2\,s_h^2}{n_h} \tag{13}$$

where s_h^2 is the sample variance from the hth zone.

The efficiency of stratified sampling in estimating mean values can be compared to that of random sampling by the following relationship using Eqs. (2) and (12).

$$\text{Relative Efficiency} = \frac{\mathrm{Var}\,(\overline{X}\ \text{by random sampling})}{\mathrm{Var}\,(\overline{X}\ \text{by stratified sampling})}$$

$$= \frac{\sigma^2/n}{\displaystyle\sum_{h=1}^{L} \frac{W_h^2\,\sigma_h^2}{n_h}} \tag{14}$$

It can be seen from Eq. (14) that stratified sampling efficiency increases for either an increase in σ^2 or a decrease in σ_h^2. Therefore, in order to achieve high efficiency using this method, the structure should be zoned so that the statistical properties of the zones are quite different from each other but the data within each zone are very similar. In practice, the within zone variances will be found to be smaller than the overall structure

variance, so stratified sampling will generally be superior to random sampling. This may be true even though the zoning operation is performed after an initial attempt at random sampling. Often the results of random sampling will provide the best indicators for natural zone boundaries. For example, assume that mean square stress has been measured at nine points selected at random on an aircraft wing as illustrated in Figure 1.

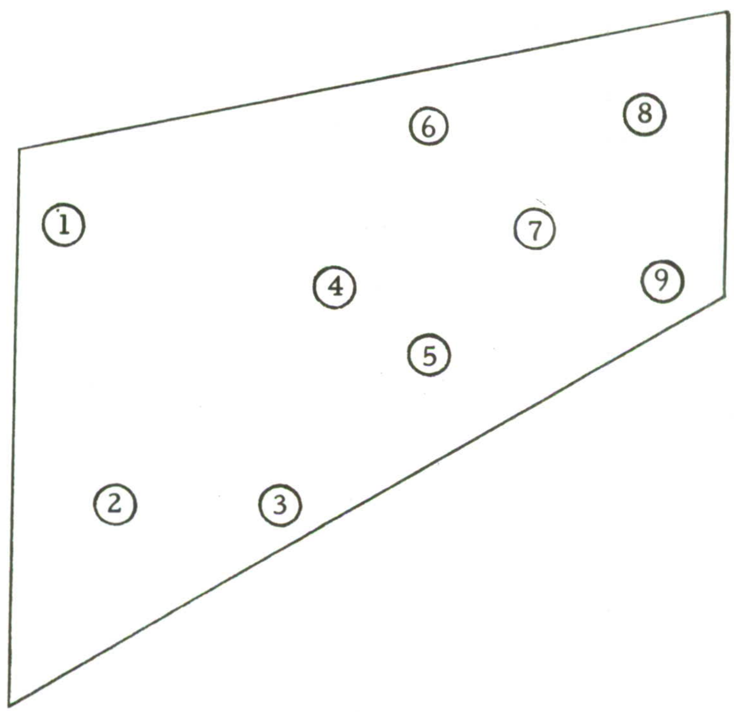

Figure 1

Further, assume the set of data values at each point given in Table 1 has been obtained from long records during stationary loading conditions so that the statistical uncertainty of each measurement is negligible.

Data Point	Measured ms Stress X_i, $(\text{psi})^2$
1	10.2×10^7
2	$11.5 \times$
3	$5.6 \times$
4	$7.1 \times$
5	$5.9 \times$
6	$6.3 \times$
7	$8.7 \times$
8	$9.5 \times$
9	$9.0 \times$

Table 1

Then, if it is required to estimate the mean value of mean square stress in the wing with low variance in the estimate, proceed as follows.

Using Eq. (1) to compute the sample mean for the values in Table 1 results in

$$\overline{X} = \frac{1}{9} \sum_{i=1}^{9} X_i = 8.2 \times 10^7 \ (\text{psi})^2$$

The sample variance is, from Eq. (4),

$$s_X^2 = \frac{1}{8}\sum_{i=1}^{9}(X_i - 8.2 \times 10^7)^2 = 4.3 \times 10^{14} (\text{psi})^4$$

and the estimated variance of the sample mean is given by Eq. (3) for random sampling.

$$\widehat{\text{Var}}\,(\bar{X}) = \frac{4.3 \times 10^{14}}{9} = .48 \times 10^{14} (\text{psi})^4$$

Now, noting the relationship between the stress level and the physical location of the various points in the sample, it would seem likely that greater sampling efficiency (less variance in the estimate of the mean) could be achieved if the wing were partitioned into zones such as shown in Figure 2.

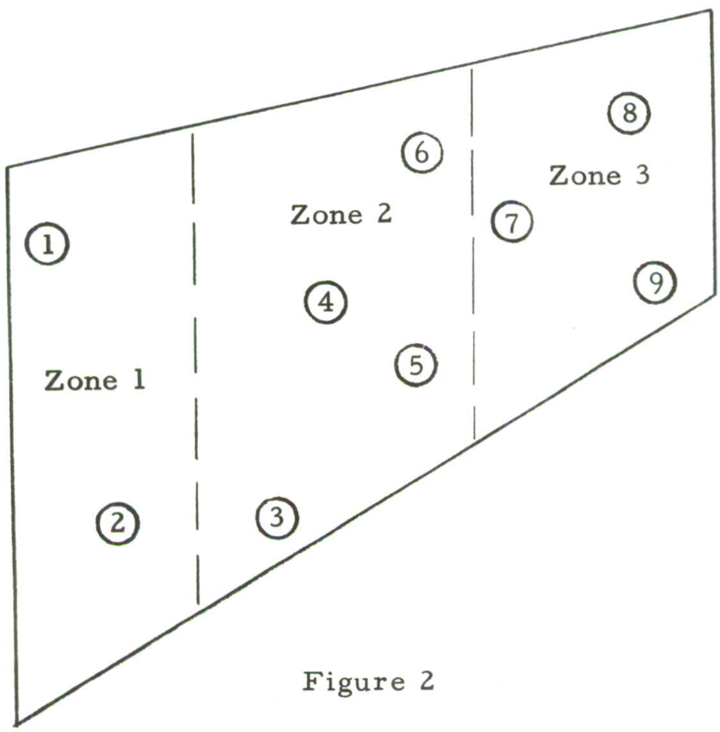

Figure 2

Examining the same data from the standpoint of stratified sampling, where the weighting factor W_h is based on the ratio of zone area to wing area alone, the parameters can be summarized as in Table 2.

Zone	W_h	n_h	X_i
1	.33	2	10.2×10^7
			$11.5 \times$
2	.37	4	$5.6 \times$
			$7.1 \times$
			$5.9 \times$
			$6.3 \times$
3	.30	3	$8.7 \times$
			$9.5 \times$
			$9.0 \times$

Table 2

The sample mean calculated using Eq. (10) is

$$\bar{X} = \sum_{h=1}^{3} W_h \bar{X}_h = 8.7 \times 10^7 (\text{psi})^2$$

and the estimate of the variance is, by Eq. (13)

$$\widehat{\text{Var}}\,(\bar{X}) = \sum_{h=1}^{3} \frac{W_h^2 \, s_h^2}{n_h} = 0.066 \times 10^{14} (\text{psi})^4$$

Thus, it can be concluded that, although the estimate of the mean value of mean square stress in the wing is nearly the same when using random or stratified sampling, the latter method gives much more confidence in the result.

The probability that the mean square stress measured at a point in any of the zones will equal or exceed a specified critical level can be estimated using a technique similar to that applied in Section 2.1 for random sampling. That is, a random variable Y assumes the value Y_i equal to one if $X_i \geq X_c$ and zero if $X_i < X_c$. Then the desired probability can be estimated from

$$\hat{P} = \sum_{h=1}^{L} W_h \hat{P}_h \tag{15}$$

where

$$\hat{P}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_i \tag{16}$$

The variance is then estimated by

$$\widehat{\mathrm{Var}}\,(\hat{P}) = \sum_{h=1}^{L} W_h^2 \frac{\hat{P}_h(1 - \hat{P}_h)}{n_h - 1} \tag{17}$$

3. OPTIMUM ALLOCATION IN STRATIFIED SAMPLING

Assuming that stratified sampling has been chosen as the tool for estimating statistical properties of mean square stress in a structure, the optimum allocation of a fixed sample size n can be determined from the following. For the case where the mean value of spatially distributed mean square stress measurements is to be estimated, the optimum allocation is obtained by minimizing Eq. (12) with respect to n_h. Letting the weighting function depend on area ratios alone, the allocation for each zone is determined from

$$n_h = n \frac{A_h \sigma_h}{\sum_{h=1}^{L} A_h \sigma_h} \tag{18}$$

As an example of the use of Eq. (18), suppose the problem is to obtain the best estimate of the mean value of mean square stress in the wing illustrated in Figures 1 and 2 employing a total of nine transducers and making use of the preliminary data in Tables 1 and 2. After zoning the wing structure as shown in Figure 2 and computing the area and sample variance for each zone, optimum allocation is determined below in Table 3.

Zone	A_h	s_h	$A_h s_h$	$A_h s_h / \sum A_h s_h$	n_h
1	44	$.92 \times 10^7$	40.5×10^7	.46	4
2	49	.65x	31.8 x	.35	3
3	40	.41 x	16.4 x	.19	2

Table 3

As would be expected, optimum allocation requires a larger proportion of the measurements to be assigned to the larger and more variable zones. In the special case when the variances σ_h^2 are equal for all zones, the optimum allocation of n reduces to

$$n_h = n \frac{A_h}{A} \tag{19}$$

That is, the allocation for each zone should be proportional to structural area alone (or other significant weighting factor considerations). This special case is called "simple stratification" and is the most commonly used method when the zone variances are unknown.

When sampling is conducted for the purpose of determining the probability that a mean square stress measured at any point exceeds a critical value, the optimum allocation of sample points is given by

$$n_h = n \frac{A_h\sqrt{P_h(1 - P_h)}}{\sum_{h=1}^{L} A_h\sqrt{P_h(1 - P_h)}} \tag{20}$$

The parameter P_h is, of course, unknown and must be either assumed or estimated from preliminary data.

If the object of the investigation is to make comparisons between different zones, the rules for allocating the number of samples to each zone are slightly different from those which applied for the above developments. For example, it may be desired to compare the mean of the measurements of two regions of a structure, two different structures, or similar structures on two aircraft. If $\overline{X}_1$ and $\overline{X}_2$ denote the means of the data in two regions of interest,

$$\text{Var}\,(\overline{X}_1 - \overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{21}$$

If $\text{Var}\,(\overline{X}_1 - \overline{X}_2)$ is minimized with respect to n_1 and n_2, one obtains

$$n_h = n \frac{\sigma_h}{\sigma_1 + \sigma_2} \quad , \quad h = 1, 2 \tag{22}$$

Equation (22) indicates that the number of samples in each region should be allocated proportionally to the standard deviation. Note that the sample allocation given by Eq. (22) is independent of the size of the area being considered, while that given previously by Eq. (18) was directly proportional to the area of interest. In general, if there are L regions of comparative interest, the optimum allocation among them would be (see Reference 2)

$$n_h \approx n \frac{\sigma_h}{\sum_{h=1}^{L} \sigma_h} \tag{23}$$

4. DETERMINATION OF SAMPLE SIZE

In estimating the statistical properties of a random process from a sample consisting of a finite number of data points, the accuracy of the estimates increases with sample size. Although a high degree of accuracy is always desirable, economic considerations usually impose a practical restriction upon the maximum number of points in a sample. Therefore, practical sampling procedures involve a compromise between accuracy and economy. The logical first step, then, is to define the amount of error that can be tolerated in the sample estimates. This can take the form of a statement of precision specifying the minimum probability that the difference between an estimate and the corresponding true value does not exceed a given amount. Then, if sufficient information about the distribution of the variable under investigation is available, a rational approach to the determination of the sample size for a given experiment can be implemented. In this section the sample size requirements associated with two specific, well-known distributions are discussed as well as a nonparametric approach to sample size determination.

4.1 SAMPLE SIZE UNDER A NORMALITY ASSUMPTION

Assume that the distribution of sample mean square stress measurements, X, in a structure tends to normality as the sample size increases. Although this assumption suffers under the practical limitation that mean square stress can never be negative, this should not materially decrease the value of the following developments.

Consider the problem of estimating the mean value, $\overline{X}$, of the data within plus or minus d units of the true mean μ. The appropriate statement of precision for a probability of $(1 - \alpha)$ is expressed as

$$\text{Prob}\left(| \overline{X} - \mu | \leq d \right) \geq 1 - \alpha \tag{24}$$

Now let

$$Z_{\alpha/2} = \frac{\overline{X} - \mu}{SE(\overline{X})} \tag{25}$$

where $Z_{\alpha/2}$ is the 100 $\alpha/2$ percentage point of the standardized normal distribution, and

$$SE(\overline{X}) = \frac{\sigma}{\sqrt{n}} \tag{26}$$

is the standard error in estimating mean values obtained from Eq. (2). A value for σ must either be assumed or estimated from preliminary measurements. Combining Eqs. (25) and (26), the sample size required to satisfy Eq. (24) is

$$n = \frac{\sigma^2 Z^2_{\alpha/2}}{d^2} \tag{27}$$

To illustrate the approach, consider the following example. Suppose it is required to estimate the mean value of mean square stress in the wing shown in Figure 1 within $\pm 10^7$ $(psi)^2$ with a minimum probability of .90 under a normality assumption. That is, if the experiment were repeated many times, the sample mean would fall between $\mu + 10^7$ and $\mu - 10^7$ $(psi)^2$ on at least 90% of the trials. In this case, $d = 10^7$, $1 - \alpha = .90$, $Z_{.05} = 1.65$, and σ^2 is taken as equal to the sample variance from Section 2.3, or $\sigma^2 = 4.3 \times 10^{14}$ $(psi)^4$. Then,

$$n = \frac{4.3 \times 10^{14} \; (1.65)^2}{10^{14}} = 11.7$$

Therefore, a sample size of 12 would be appropriate for the required accuracy. Although the tolerance $\pm 10^7$ $(psi)^2$ might seem large, in this case it equals only about $\pm$ 12% of the first estimate for the mean computed in the example of Section 2.3.

The problem of estimating the probability that the stress at a point selected on the structure at random will have a mean square value, X,

exceeding an established critical level X_c, is approached in a similar manner. If it is required that the estimated probability be within plus or minus d units of the true probability, the statement of precision is

$$\text{Prob}\left(\left|\hat{P} - P\right| \leq d\right) \geq 1 - \alpha \tag{28}$$

where P is defined as

$$P = \frac{1}{\sigma\sqrt{2\pi}} \int_{X_c}^{\infty} e^{-(x-\mu)^2/2\sigma^2} \, dx \tag{29}$$

If the sample size is large (say $n > 10$), the function

$$Z = \frac{\hat{P} - P}{SE(\hat{P})} \tag{30}$$

is approximately normal with zero mean and unit variance. Since the standard error in estimating P is [see Eq. (8)],

$$SE(\hat{P}) = \sqrt{\frac{P(1 - P)}{n - 1}} \tag{31}$$

the required sample size in this case is given by

$$n = \frac{Z^2_{\alpha/2} P(1 - P)}{d^2} + 1 \tag{32}$$

4.2 SAMPLE SIZE UNDER A LOG NORMAL ASSUMPTION

Since the mean square value of a stress can never be negative, the normality assumption could clearly not hold true for all data. Some form of skewed distribution might then provide a more suitable model for this or similar situations. One which has been well studied and which will be described in this section is the log normal distribution.

The random variable X is said to have a log normal distribution if $Y = \log X$ is normally distributed. That is, if Y is normal, then $X = e^{Y}$ is log normal.

The density function of X, $f_X(x)$ is

$$f_X(x) = \frac{d}{dx} P(X \leq x) = \frac{d}{dx} F_Y(\log x) = \frac{1}{x} f_Y(\log x)$$

$$= \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} \quad \text{for } x \geq 0 \tag{33}$$

$$= 0 \quad \text{for } x < 0$$

where F_Y denotes a normal distribution function with mean μ and variance σ^2. The mean of X, μ_X, in terms of the parameters of Y is

$$\mu_X = \int_0^{\infty} x \, f_X(x) \, dx$$

$$= \int_{-\infty}^{\infty} e^y \, f_Y(y) \, dy \tag{34}$$

$$= e^{\mu} + \sigma^2/2$$

The variance of X, σ_X^2, is

$$\sigma_X^2 = E(X^2) - \mu_X^2$$

$$= \int_0^{\infty} x^2 \, f_X(x) \, dx - \mu_X^2$$

$$= \int_{-\infty}^{\infty} e^{2y} \, f_Y(x) \, dy - e^{2\mu + \sigma^2} \tag{35}$$

$$= e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2}$$

$$= e^{2\mu + \sigma^2} \left(e^{\sigma^2} - 1 \right)$$

It is important to note that the transformation, e^Y, transforms the mean as well as the variance.

The maximum likelihood estimates of μ_X and σ_X^2 can be shown to be

$$\hat{\mu}_X = e^{\overline{Y}+s^2/2} \tag{36}$$

and

$$\hat{\sigma}_X^2 = e^{2\overline{Y}+s^2}\left(e^{s^2} - 1\right) \tag{37}$$

where $\overline{Y}$ and s^2 denote the sample mean and variance of the normal distribution log X.

If Eqs. (34) and (35) are solved for μ and σ^2,

$$\mu = \log \mu_X - \frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) = \log \frac{\mu_X}{\sqrt{1 + \frac{\sigma_X^2}{\mu_X^2}}} \tag{38}$$

and

$$\sigma^2 = \log\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \tag{39}$$

Let

$$Z = \frac{\overline{\log X} - \mu}{\sigma / \sqrt{n}} \tag{40}$$

where

$$\overline{\log X} = \frac{\sum_{i=1}^{n} \log X_i}{n} \tag{41}$$

and σ^2 is estimated by

$$s^2 = \frac{\sum_{i=1}^{n} (\log X_i - \overline{\log X})^2}{n - 1} \tag{42}$$

The sample size n which is required to put $\overline{\log X}$ within plus or minus d' units of μ with a probability of $(1 - \alpha)$ is, from Eq. (27),

$$n = \frac{\sigma^2 Z^2_{\alpha/2}}{(d')^2} \tag{43}$$

where $Z_{\alpha/2}$ denotes the 100 $\alpha/2$ percentage point of the standardized normal distribution. That is, if $\overline{\log X}$ is computed from n samples, then

$$\text{Prob}\left(\left| \log \overline{X} - \mu \right| \leq d'\right) \geq 1 - \alpha \tag{44}$$

Now we will develop a formula to determine the sample size to estimate μ_X by $\overline{X}$ with a maximum error of d. Note that $\overline{\log X}$ estimates μ, but not μ_X. An estimator for μ_X is derived as follows. Let

$$w = \overline{\log X} + \frac{\sigma^2}{2} \tag{45}$$

Since it is clear from Eq. (34) that

$$\log \mu_X = \mu + \frac{\sigma^2}{2} \tag{46}$$

w estimates $\log \mu_X$, or e^w estimates μ_X.

Let n be the sample size to satisfy the following relationship:

$$\text{Prob}\left(\left| w - \log \mu_X \right| \leq d' \right) \geq 1 - \alpha \tag{47}$$

substituting Eqs. (45) and (46) into (47),

$$\text{Prob}\left(\left| \overline{\log X} - \mu \right| \leq d' \right) \geq 1 - \alpha$$

In other words, Eqs. (44) and (47) are equivalent, and Eq. (43) can be used to estimate the required sample size to satisfy Eq. (47). Equation (47) can be written as

$$\text{Prob}\left(w - d' < \log \mu_X < w + d' \right) \geq 1 - \alpha \tag{48}$$

or

$$\text{Prob}\left(e^{w-d'} < \mu_X < e^{w+d'} \right) \geq 1 - \alpha$$

Now, the objective is to determine the sample size n such that $\bar{X}$ will satisfy the following.

$$\text{Prob}\left(\bar{X} - d < \mu_X < \bar{X} + d \right) \geq 1 - \alpha \tag{49}$$

By comparison with Eq. (48),

$$X - d = e^{w-d'} \tag{50}$$

or

$$X + d = e^{w+d'} \tag{51}$$

Denoting solutions for d' in Eqs. (50) and (51) by d'_1 and d'_2, respectively, it follows that

$$d'_1 = w - \log(\overline{X} - d) \tag{52}$$

$$d'_2 = \log(\overline{X} + d) - w \tag{53}$$

Let n_1 and n_2 be the sample sizes obtained by substituting d'_1 and d'_2 into Eq. (43). If n is defined by

$$n = \max(n_1, n_2) \tag{54}$$

then n is the sample size sufficient to assure Eq. (49). Summarizing, the sample size n which is required to put $\overline{X}$ within $\pm d$ units of μ_X with a probability of $1 - \alpha$ is given by

$$n = \max\left\{ \frac{\sigma^2 Z^2_{\alpha/2}}{\left[\overline{\log X} + \frac{\sigma^2}{2} - \log(\overline{X} - d)\right]^2} ; \frac{\sigma^2 Z^2_{\alpha/2}}{\left[\log(\overline{X} + d) - \overline{\log X} - \frac{\sigma^2}{2}\right]^2} \right\} \tag{55}$$

For example, suppose the illustrative example of Section 4.1 is reworked using a log normal assumption for the sampling distribution. The parameters required for a solution by Eq. (55) include $\overline{X}$ and $\overline{\log X}$ which can be computed from the preliminary data given in Section 2.3 and σ^2 which can be estimated by s^2.

Assume that μ_X is to be estimated with an error d of less than $10^7 (\text{psi})^2$ with a probability $1 - \alpha$ of .90. Then, from the data in Table 1, and from Eqs. (41) and (42),

$$\overline{\log X} = \frac{1}{9} \sum_{i=1}^{9} \log X_i = 18.1984$$

$$s^2 = \frac{1}{8} \sum_{i=1}^{9} (\log X_i - 18.1984)^2$$

$$= .0697$$

and

$$\log(\overline{X} \pm d) = \log(8.2 \times 10^7 \pm 10^7)$$

$$= 18.3373 \; ; \; 18.0922$$

Then,

$$n = \max \left\{ \frac{.0697\,(1.65)^2}{\left(18.1984 + \frac{.0697}{2} - 18.0922\right)^2} \; ; \; \frac{.0697\,(1.65)^2}{\left(18.3373 - 18.1984 - \frac{.0697}{2}\right)^2} \right\}$$

$$= \max \left\{ 0.5 \; ; \; 17.5 \right\}$$

Therefore, the sample size required to assure the stated accuracy under a log normal assumption would be 18. Note that since the assumed distribution is skewed, the estimation error is not symmetrical about the mean.

Sample size requirements associated with estimating the probability of exceeding a critical level are determined as they were for the normal distribution. That is, Eq. (32) applies for the log normal case. However, instead of defining P as in Eq. (29), P in this case is the integral of Eq. (33).

$$P = \frac{1}{\sigma\sqrt{2\pi}} \int_{X_c}^{\infty} \frac{1}{x}\, e^{-(\log x - \mu)^2/2\sigma^2}\, dx \tag{56}$$

4.3 A NONPARAMETRIC METHOD FOR SAMPLE SIZE DETERMINATION

For the case when it is judged undesirable to assume a specific distribution of data values, the Tchebycheff inequality can be applied to determine a sample size nonparametrically. The Tchebycheff inequality states that for every k,

$$\text{Prob}\left(\left| \overline{X} - \mu \right| \leq k \frac{\sigma}{\sqrt{n}} \right) \geq 1 - \frac{1}{k^2} \tag{57}$$

The significance of this relationship is that the area under a probability density curve located outside of $\mu \pm k\sigma$ will not exceed $1/k^2$ regardless of the distribution. Using the notation of Eq. (24),

$$\alpha = \frac{1}{k^2} \qquad \text{or} \qquad k = \frac{1}{\sqrt{\alpha}}$$

$$d = k \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{\alpha n}} \tag{58}$$

Thus, a conservative estimate of the number of points in a sample required to place $\overline{X}$ within $\pm d$ units of the true mean μ with probability $(1 - \alpha)$ for any distribution of X is

$$n = \frac{\sigma^2}{\alpha d^2} \tag{59}$$

Similarly, the sample size required to estimate P within a given tolerance for a specified probability is given by

$$n = \frac{P(1 - P)}{\alpha d^2} \tag{60}$$

It can easily be shown that for any P in the interval zero to one, $P(1 - P) \leq 1/4$. Therefore, the upper bound on Eq. (60) is

$$n = \frac{1}{4\alpha d^2} \tag{61}$$

To demonstrate the fact that this method leads to conservative sample size requirements, consider the application of Eq. (59) to the previous example problem. Using the preliminary data in Table 1 to compute an estimate for σ^2, and specifying an allowable error of $\pm 10^7$ $(psi)^2$ with probability $1 - \alpha = .90$, the total number of measurement points would be

$$n = \frac{4.3 \times 10^{14}}{.10\ (10^7)^2} = 43$$

Clearly, this nonparametric method is quite inefficient when additional information about the sampling distribution exists. However, it does represent a bound on the sample size and has engineering applications.

REFERENCES

1. Choi, S. C., and L. D. Enochson, "Random Fatigue Test Sampling Requirements," AFFDL-TR-65-95, AF Flight Dynamics Laboratory, Wright-Patterson AFB, Ohio. August 1965.

2. Cochran, W. G., Sampling Techniques, John Wiley and Sons, Inc. New York, 1963.

APPENDIX

SUMMARY OF IMPORTANT RELATIONSHIPS

1. RANDOM SAMPLING

a. Mean

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Eq. (1)
page 2

b. Variance

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Eq. (4)
page 3

c. Probability of Exceeding Critical Level X_c

$$\hat{P} = \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad Y_i = \begin{cases} 1 & \text{if} \quad X_i \geq X_c \\ 0 & \text{if} \quad X_i < X_c \end{cases}$$

Eq. (6)
page 3

2. STRATIFIED SAMPLING

a. Mean

$$\overline{X} = \sum_{h=1}^{L} W_h \overline{X}_h$$

Eq. (10)
page 5

b. Probability of Exceeding Critical Level X_c

$$\hat{P} = \sum_{h=1}^{L} W_h \hat{P}_h$$

Eq. (15)
page 11

3. OPTIMUM ALLOCATION IN STRATIFIED SAMPLING

a. Mean Value Estimation

$$n_h = \frac{A_h \sigma_h}{\sum_{h=1}^{L} A_h \sigma_h}$$

Eq. (18) page 12

$$n_h = \frac{A_h}{A} \quad \text{(simple stratification)}$$

Eq. (19) page 13

b. Probability Estimation

$$n_h = n \frac{A_h \sqrt{P_h(1 - P_h)}}{\sum_{h=1}^{L} A_h \sqrt{P_h(1 - P_h)}}$$

Eq. (20) page 13

c. Comparison of Mean Values in L Zones

$$n_h \approx n \frac{\sigma_h}{\sum_{h=1}^{L} \sigma_h}$$

Eq. (23) page 14

4. SAMPLE SIZE UNDER A NORMALITY ASSUMPTION

a. Mean Value Estimation

$$n = \frac{\sigma^2 Z^2_{\alpha/2}}{d^2}$$

Eq. (27)
page 16

b. Probability Estimation

$$n = \frac{Z^2_{\alpha/2} P(1 - P)}{d^2} + 1$$

Eq. (32)
page 17

5. SAMPLE SIZE UNDER A LOG NORMAL ASSUMPTION

a. Mean Value Estimation

$$n = \max \left\{ \frac{\sigma^2 Z^2_{\alpha/2}}{\left[\overline{\log X} + \frac{\sigma^2}{2} - \log(\overline{X} - d) \right]^2} ; \frac{\sigma^2 Z^2_{\alpha/2}}{\left[\log(\overline{X} + d) - \overline{\log X} - \frac{\sigma^2}{2} \right]^2} \right\}$$

Eq. (55)
page 24

b. Probability Estimation

$$n = \frac{Z^2_{\alpha/2} P(1 - P)}{d^2} + 1$$

Eq. (32)
page 17

6. SAMPLE SIZE NONPARAMETRICALLY

a. Mean Value Estimation

$$n = \frac{\sigma^2}{\alpha d^2}$$

Eq. (59)
page 26

b. Probability Estimation

$$n = \frac{P(1 - P)}{\alpha d^2}$$

Eq. (60)
page 27

$$\leq \frac{1}{4\alpha d^2}$$

Eq. (61)
page 27

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Measurement Analysis Corporation, 10960 Santa Monica Boulevard, Los Angeles, California 90025	2a. REPORT SECURITY CLASSIFICATION: Unclassified 2b. GROUP: N/A

3. REPORT TITLE

Design of Sampling Procedures for Flight Loads Studies

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Final Report (one of two under this contract)

5. AUTHOR(S) (Last name, first name, initial)

Choi, S. C.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May 1967	38	2

8a. CONTRACT OR GRANT NO. AF 33(615)-3123	9a. ORIGINATOR'S REPORT NUMBER(S) AFFDL-TR-66-202
b. PROJECT NO. BPSN 5(61136716-62405334)	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	MAC 510-03

10. AVAILABILITY/LIMITATION NOTICES This document is subject to special export controls and each transmittal to foreign governments or foreign nationals may be made only with prior approval of the Air Force Flight Dynamics Laboratory (FDTR), Wright-Patterson Air Force Base, Ohio 45433.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
None	Air Force Flight Dynamics Laboratory (FDTR) Wright-Patterson AFB, Ohio 45433

13. ABSTRACT

The problems of estimating the total number of measurement points and the optimum spatial distribution of locations on a structure are approached theoretically in this report. The significant factors to be considered are statistical reliability and economy. Therefore, the relationships are developed with the emphasis on measurement efficiency. Random, systematic, and stratified sampling methods are compared for efficiency in estimating mean values. Then the optimum allocation of a fixed number of measurement points in stratified sampling is developed, and illustrative examples are given. Finally, relationships are presented which will allow the total sample size to be estimated under the assumptions of normal and log-normal sampling distributions as well as by a nonparametric approach. These formulas are deemed to be quite useful for experiment planning purposes. This abstract is subject to special export controls and each transmittal to foreign governments or foreign nationals may be made only with prior approval of the Air Force Flight Dynamics Laboratory (FDTR), Wright-Patterson Air Force Base, Ohio 45433.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
Optimum Sampling Flight Loads Data Random Processes Structural Zoning						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through ______________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through ______________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through ______________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.